A stand-off XML-TEI representation of reference annotation

Aria Adli, Eric Engel, Laurent Romary, Fahime Same

A stand-off XML-TEI representation of reference annotation

Aria Adli[1], Eric Engel[1], Laurent Romary[2], Fahime Same[1]

[1]University of Cologne, [2]INRIA
{aria.adli, eric.engel, f.same}@uni-koeln.de, laurent.romary@inria.fr

Background and previous work

- Multilingual corpus project on French, Spanish, Persian, and Catalan
- Spontaneous speech data elicited in a game task
- More info: `www.sgscorpus.com`

- Building on extensive work on (stand-off) representation of referring expressions and referential links [4,6] and types of links [5]
- So far, no existing models making a clean distinction between expression-related and entity-related features and links

What's new?

- Reference annotation is an essential element in corpus-linguistic research at the interface of syntax and discourse structure.
- We propose a representation based on a semantically plausible model distinguishing referring expressions, discourse referents, and links.
  - Neater representation of grammatical *vs.* semantic features and of phenomena of indirect reference ("bridging")
  - More efficient annotation of longer stretches of natural discourse
  - Integrated in a multilayer stand-off annotation in compliance with TEI-ISO standards: MAF [1], SynAF [2], RAF [3]

Example

(Persian, Interview 12, line 121):

(1) dAdAS-eS-o doxtar-eS ke az SahrestAn umade budan.
brother-3SG.POSS-and daughter-3SG.POSS C-TOP from province came
'His brother and his daughter came from the province.'

Links

```
<so:standOff type="reference">

    <so:listAnnotation type="link">
        <linkGrp type="objectalLink">
            <link xml:id="link1" target="#ent1 #ent4" ana="#link1−ana"/>
            <link xml:id="link2" target="#ent3 #ent4" ana="#link2−ana"/>
        </linkGrp>
        <fs xml:id="link1−ana"><symbol value="subset"/></fs>
        <fs xml:id="link2−ana"><symbol value="subset"/></fs>
    </so:listAnnotation>
```

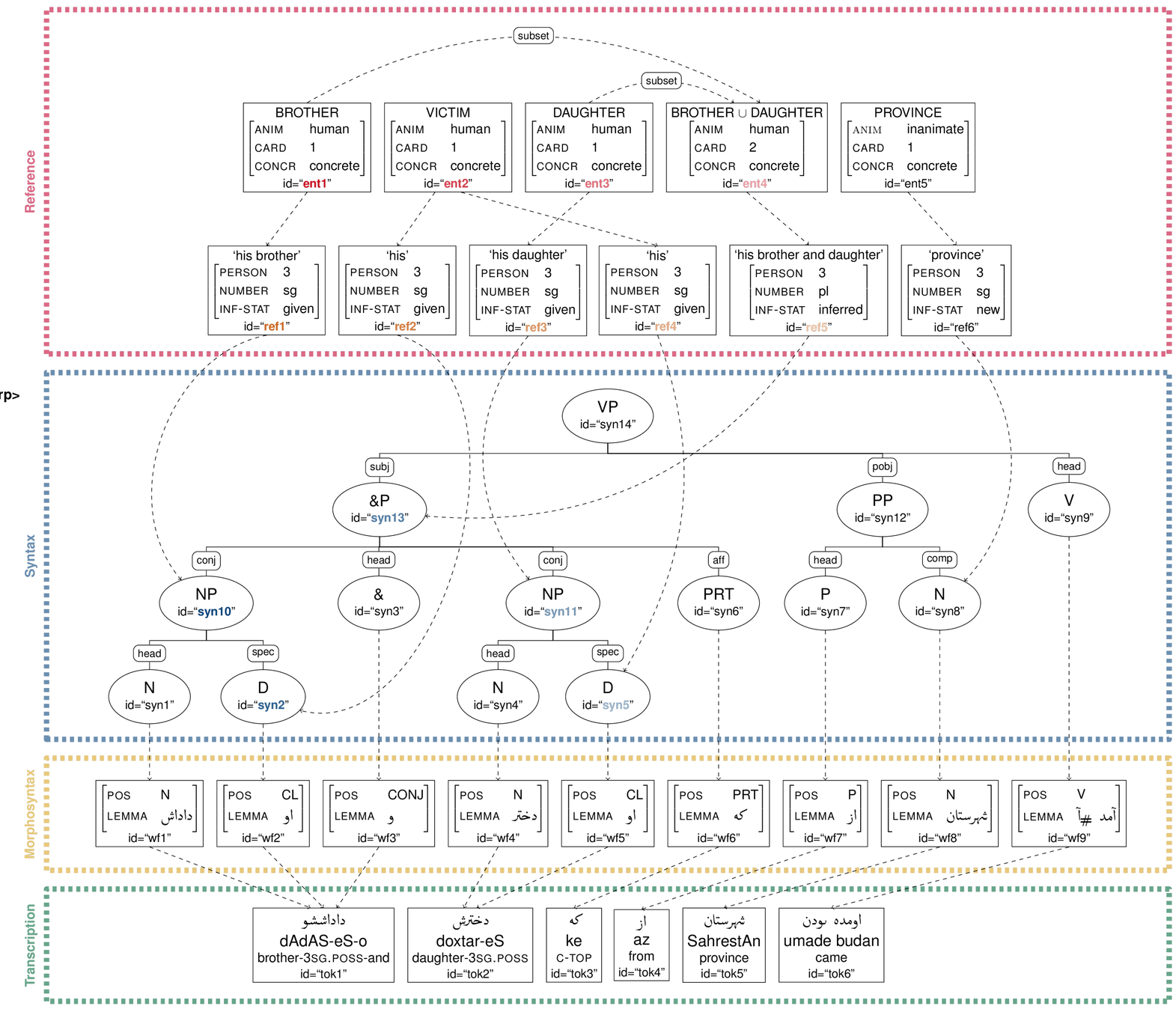
Discourse referents

```
    <so:listAnnotation type="discourseEntity">
        <interpGrp>
            <interp xml:id="ent1" inst="#ref1" ana="ent1−ana">brother_of_victim</interp>
            <interp xml:id="ent2" inst="#ref2 #ref4" ana="ent2−ana">victim</interp>
            <interp xml:id="ent3" inst="#ref3" ana="ent3−ana">daughter_of_victim</interp>
            <interp xml:id="ent4" inst="#ref5" ana="ent4−ana">brother_and_daughter</interp>
        </interpGrp>
        <fs xml:id="ent1−ana">
            <f name="animacy"><symbol value="human"/></f>
            <f name="abstractness"><symbol value="concrete"/></f>
        </fs>
        <!−− ... −−>
        <fs xml:id="ent4−ana">
            <f name="animacy"><symbol value="human"/></f>
            <f name="abstractness"><symbol value="concrete"/></f>
        </fs>
    </so:listAnnotation>
```

Referring expressions

```
    <so:listAnnotation type="referringExpression">
        <annotationBlock xml:id="ref−u−12−121" target="#u−12−121">
            <spanGrp>
                <span xml:id="ref1" target="#syn10" ana="#ref1−ana"/>
                <span xml:id="ref2" target="#syn2" ana="#ref2−ana"/>
                <span xml:id="ref3" target="#syn11" ana="#ref3−ana"/>
                <span xml:id="ref4" target="#syn5" ana="#ref4−ana"/>
                <span xml:id="ref5" target="#syn13" ana="#ref5−ana"/>
            </spanGrp>
            <fs xml:id="ref1−ana">
                <f name="person"><symbol value="3"/></f>
                <f name="grammaticalNumber"><symbol value="singular"/></f>
                <f name="informationStatus"><symbol value="given"/></f>
            </fs>
            <!−− ... −−>
            <fs xml:id="ref5−ana">
                <f name="person"><symbol value="3"/></f>
                <f name="grammaticalNumber"><symbol value="plural"/></f>
                <f name="informationStatus"><symbol value="inferred"/></f>
            </fs>
        </annotationBlock>
    </so:listAnnotation>

</so:standOff>
```



Model features

- Separation of primary data (transcription) and secondary data (annotations)
- Incremental build-up of annotation layers
- Three-way distinction of reference-related features
  - Local **grammatical features** (grammatical gender, number, person) and sentence-bound discourse features (information status) apply to **referring expressions**.
  - Inherent **semantic features** of referents (animacy, cardinality, specificity, concreteness/abstractness) apply to the **referents**. Their values are constant over the entire discourse.
  - **Relations between referents** (subset, partOf) are recognized as **permanent**, allowing for a better representation of so-called bridging phenomena in discourse.

Future work

- User-friendly visualization and query interface (in ANNIS)
- Integration of different levels of information structure analysis
  - [±topical] as feature of a referring expression?
  - Markup of focus domains directly on syntactic phrases?
- Further expansion of the multilayer architecture to the analysis of other cross-sentential phenomena, e.g., rhetorical structure, dialogue analysis

References

[1] ISO 24611:2012 Language resource management — Morpho-syntactic annotation framework (MAF).
[2] ISO 24615-1:2014 Language resource management — Syntactic annotation framework (SynAF) — Part 1: Syntactic model.
[3] ISO/AWI 24617-9 [Under development] Language resource management — Semantic annotation framework — Part 9: Reference Annotation Framework.
[4] Poesio, M., 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 154-162.
[5] Riester, A. and Baumann, S., 2017. The RefLex scheme - Annotation guidelines. *SinSpeC* 14. 1-31.
[6] Salmon-Alt, S. and Romary, L. 2005. The Reference Annotation Framework: A case for semantic content representation. In H. Bunt (ed.), *IWCS-6*, Tilburg, Netherlands: ACL SIGSEM.